DeepSeek与AI幻觉

清华大学

新闻与传播学院新媒体研究中心

人工智能学院

双聘教授 @新媒沈阳 团队

博士后　张家铖

2025年2月

案例1

Prompt：给我一些DeepSeek在金融行业的应用案例

DeepSeek：

- 某头部银行利用DeepSeek构建因果归因网络，识别小微企业违约的隐性因素(如上下游账期错配)，针对性设计“供应链票据贴现+账期保险”组合产品，不良率下降4.2个百分点。
- 国信证券在“金太阳APP”中部署DeepSeek-R1-Distill-32B端侧模型，客户本地输入风险测评数据后，模型生成个性化投资组合建议，仅将匿名化策略权重同步至云端风控系统，数据泄露风险降低90%。

国产大模型DeepSeek成券商新宠！本地化部署加速金融科技迭代！

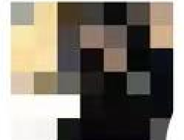
金融界 2025-02-10 07:53

近日，国产大模型DeepSeek在金融业掀起了一股热潮，又有近10家券商官宣完成DeepSeek-R1的本地化部署。至此，证券业探索DeepSeek模型应用的券商至少已有16家。DeepSeek-R1作为DeepSeek系列模型的最新版本，正以其高性能、低成本的特性，迅速成为券商数字化转型的新宠。

多家券商快速行动，DeepSeek成金融科技新利器

此前，中泰证券就运用DeepSeek-V3模型，在金融新闻文本挖掘与分析领域实现突破。通过将金融专家规则融入提示词，模型能够自动绘制产业链图谱，清晰呈现实体之间的复杂关系网络。同时，模型还能精准判断新闻中提及实体的正负面情绪，准确率高达90%。

案例2



今天被Deepseek耍得团团转，给我推荐了一个阿布扎比的本地市场在迪拜，道歉后，给推荐的一个特色商场根本不存在。好，给他第三次机会，推荐一个直面滨海大道与海岸线的咖啡馆，实际是这样的😭😭

收起





案例3

Whisper：OpenAI的自动语音识别（ASR）系统

行业应用：医疗系统中，将患者与医生的对话问诊过程音频，转写为文字病例，有超过30000名临床医生和40个医疗系统使用

发现：100多个小时的Whisper转录样本，其中约有一半内容存在幻觉

原音频：“嗯，她的父亲再婚后不久就去世了”

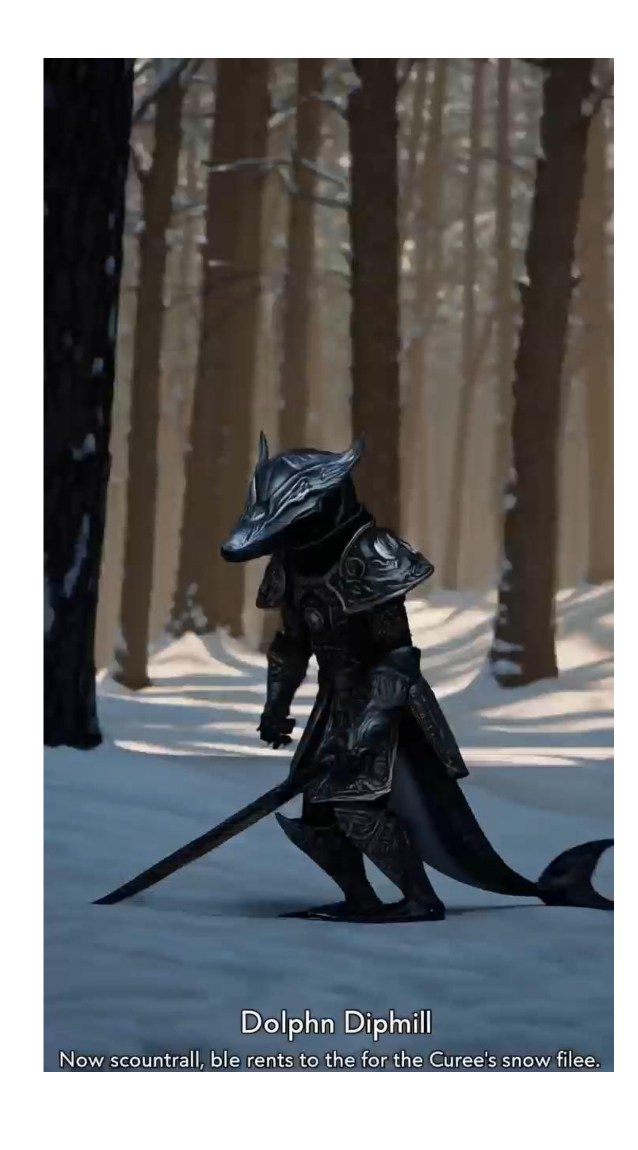
转录文本：“没关系。只是太敏感了，不方便透露。她确实在65岁时去世了”

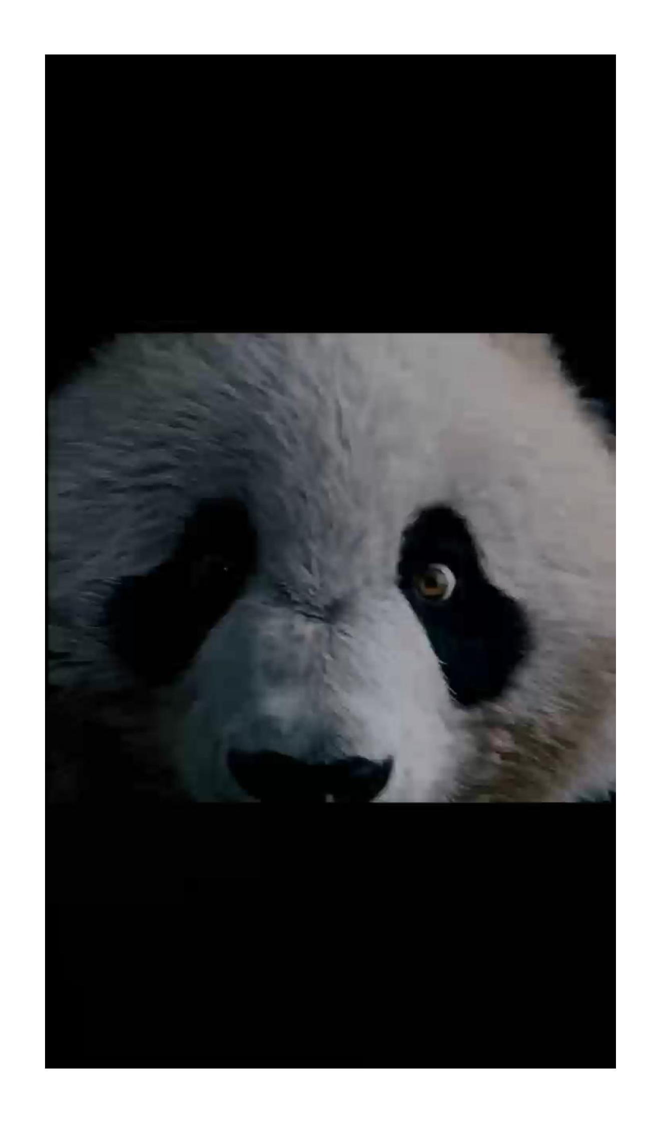
结果：2.6W多份自动转录病例中，几乎每本都存在瞎编和幻觉问题，对患者健康和医疗系统产生严重负面影响

案例4



案例5


Dolphn Diphill
Now scountrall, ble rents to the for the Curee's snow filee.



什么是AI幻觉？

学术：指模型生成与事实不符、逻辑断裂或脱离上下文的内容，本质是统计概率驱动的“合理猜测”

说人话：一本正经地胡说八道

事实性幻觉：指模型生成的内容与可验证的现实世界事实不一致

忠实性幻觉：指模型生成的内容与用户的指令或上下文不一致

提问：糖尿病患者可以通过吃蜂蜜代替糖吗？

	回答	分析
事实性幻觉	是的，蜂蜜是天然的，可以帮助糖尿病患者稳定血糖水平。	**错误：**蜂蜜虽然是天然食品，但仍然含有大量果糖和葡萄糖，会升高血糖水平，不适合糖尿病患者代替糖使用。
忠实性幻觉	蜂蜜富含维生素和矿物质，对提高免疫力很有帮助，因此是一种健康的食品。	**偏题：**回答内容虽无事实错误，但与提问“糖尿病患者是否可以用蜂蜜代替糖”无关，未忠实于用户意图。

AI为什么会产生幻觉？

- **数据偏差：**训练数据中的错误或片面性被模型放大（如医学领域过时论文导致错误结论）
- **泛化困境：**模型难以处理训练集外的复杂场景（如南极冰层融化对非洲农业的影响预测）
- **知识固化：**模型过度依赖参数化记忆，缺乏动态更新能力（如2023年后的事件完全虚构）
- **意图误解：**用户提问模糊时，模型易“自由发挥”（如“介绍深度学习”可能偏离实际需求）



Computer Science > Computation and Language

[Submitted on 22 Jan 2024]

Hallucination is Inevitable: An Innate Limitation of Large Language Models

Ziwei Xu, Sanjay Jain, Mohan Kankanhalli

音乐为什么没有幻觉？

- **音乐的主观性和多样性：**音乐是一种高度主观的艺术形式，人们对音乐的审美和理解有很大的差异。一段音乐是否“合理”或“正确”，往往取决于文化背景、个人偏好和上下文
- **音乐的抽象性：**音乐本质上是抽象的，不像文本或图像那样直接对应现实世界的具体事物。文本中的“幻觉”通常是因为模型生成的内容与事实不符，而音乐本身往往缺少明确的事实基础
- **音乐的可感知性差异：**音乐是时间性的艺术形式，即使某些部分听起来不协调或不符合预期，它们也可能在整个作品的上下文中变得合理。相比之下，文本或图像中的问题往往是瞬间可见的，容易引起注意
- **音乐“幻觉”的潜在表现：**逻辑断裂的歌词、结构混乱的旋律、风格混杂的编曲

AI幻觉的潜在风险

- **信息污染风险**：由于DeepSeek的低门槛和普及度高，大量AI生成内容涌入中文互联网，加剧了虚假信息传播的“雪球效应”，甚至污染下一代模型训练数据
- **信任危机**：普通用户难以辨别AI内容的真实性，可能对医疗建议、法律咨询等专业场景的可靠性产生长期怀疑
- **控制欠缺**：DeepSeek的**对齐**工作较其他闭源大模型有所欠缺，其开源特性也允许使用者随意使用，可能会成为恶意行为的工具
- **安全漏洞**：若错误信息被用于自动化系统（如金融分析、工业控制），可能引发连锁反应
- ……

AI幻觉评测

- **测试1：**随机生成100条**通用**提示语，模仿普通用户的真实使用场景，获取大模型回答后进行人工判断与标注，并进行交叉验证

大模型	幻觉率
DeepSeekV3	2%
DeepSeekR1	3%
Qianwen2.5-Max	2%
豆包	0

事实性幻觉评测

- **测试2：** 随机抽取300道**事实性幻觉**测试题，涵盖健康、科学、历史、文化、音乐等等多个领域，获取大模型回答后与正确答案比对，人工标注幻觉类型，并进行交叉验证

大模型	幻觉率
DeepSeekV3	29.67%
DeepSeekR1	22.33%
Qianwen2.5-Max	27.67%
豆包	19%

- **幻觉率初步测试结果：** DeepSeekV3>Qianwen2.5-Max>DeepSeekR1>豆包

事实性幻觉评测

➢ **常识错误**

提问	正确回答	DeepSeekV3	DeepSeekR1	Qianwen2.5-Max	豆包
"天上百鸟朝凤凰，朝凤凰，地上葵花向**太阳**，向太阳，哎，**北京**城里的**毛主席**，我们永远跟您走，幸福万年长罗！"这段歌词出自哪首歌曲	这段歌词出自藏族民歌《北京有个金太阳》。	这段歌词出自歌曲**《北京的金山上》**。	歌曲**《阿佤人民唱新歌》**（1972年），表达少数民族对领袖的拥护。	这段歌词出自歌曲**《东方红》**。	"天上百鸟朝凤凰，朝凤凰，地上葵花向太阳，向太阳，哎，北京城里的毛主席，我们永远跟您走，幸福万年长罗"出自**《北京的金山上》**。

事实性幻觉评测

➢ 逻辑陷阱

- **提问：** 为什么一向见钱眼开的小明仍然会被金钱蒙住双眼



已深度思考（用时 19 秒）

嗯，用户问的是"为什么一向见钱眼开的小明仍然会被金钱蒙住双眼?
个成语的意思。"见钱眼开"通常指看到钱就眼睛发亮，形容人贪财，
为贪图金钱而失去了判断力，做出错误决定。

用户的问题在于，既然小明本来就是个见钱眼开的人，为什么还会被
已经很贪财了，应该更懂得如何利用金钱，而不是被其蒙蔽。可能的
贪婪反而导致他失去了理性，无法正确评估风险，或者被眼前的利益

接下来，我需要验证这个解释是否合理。
骗，这时候他的贪婪让他无法冷静分析，

另外，可能要考虑"见钱眼开"和"被金钱
而导致的盲目。也就是说，小明的贪财

还需要考虑是否有其他因素，比如小明
这时候，即使他平时贪财，但缺乏应对

总结起来，可能的答案是：虽然小明平
择。或者，他的贪财性格使他在面对金

小明虽然一向见钱眼开，但过度贪婪反而导致其判断力丧失，最终被金钱蒙蔽。：

1. **贪婪的陷阱**

   见钱眼开体现对金钱的渴望，但过度贪婪会引发非理性行为。例如：

   - 追逐短期暴利时忽视风险（如参与传销、赌博）
   - 为快速获利突破道德底线（如受贿、诈骗）
   - 陷入"损失厌恶"心理，为挽回损失投入更多沉没成本

4. **机会成本的误判**

   被金钱蒙眼往往伴随着：

   - 忽视时间价值（如为加班费牺牲健康）
   - 错失人际关系等无形财富积累
   - 透支信用资本导致发展受限

这种现象本质是"金钱认知失调"：当物质追求与个人能力、社会环境产生结构性矛盾时，越执着于金钱反而越容易被其反噬。如同物理学的逃逸速度原理，当对金钱的渴望超过某个临界值，理性判断就会脱离正常轨道。

事实性幻觉评测

➢ **虚构事件**

提问	正确回答	DeepSeekV3	DeepSeekR1	Qianwen2.5-Max	豆包
水浒传中李逵为什么要大闹五台山？	李逵并没有大闹五台山，事实上，《水浒传》中大闹五台山的是花和尚鲁智深。	水浒传中李逵大闹五台山是因为他喝醉了酒，误闯了寺庙。	误以为宋江强抢民女（实为他人冒充），怒闯五台山讨说法。	李逵大闹五台山是因为他误以为鲁智深在寺庙中受欺负，于是出手相助，引发混乱。	水浒传中大闹五台山的是鲁智深，不是李逵。鲁智深大闹五台山是因为他出家后难以忍受寺里的清规戒律，又喝酒吃肉，还打伤了寺里的僧人等。

推理与幻觉的关系

- **DeepSeek V3：提问→ → 回答； DeepSeek R1：提问→思维链→回答**

- **推理增强→ → →幻觉率降低**

- **逻辑准确性与错误减少：** 推理能力强的模型能减少因逻辑错误导致的幻觉。例如，在数学问题中，模型若具备多步推理能力，更可能得出正确结论而非臆测答案

- **上下文理解与信息关联：** 强大的推理能力使模型更精准地捕捉上下文关联，避免因断章取义而生成虚构内容。例如，在问答任务中，模型能通过推理排除干扰选项，降低错误率

推理与幻觉的关系

- **Vectara： 摘要任务，DeepSeek V3幻觉率3.9% ，DeepSeek R1幻觉率14.3%**

	DeepSeek R1	DeepSeek V3
Vectara's HHEM 2.1	14.3%	3.9%
Google's FACTS w/ GPT-4o & Claude-3.5-Sonnet	4.37%	2.99%
Google's FACTS w/ GPT-4o & Gemini-1.5-Pro	3.09%	1.99%
Google's FACTS w/ Claude-3.5-Sonnet & Gemini-1.5-Pro	3.89%	2.69%

Table 1: Hallucination rates of DeepSeek R1 and V3 by various hallucination judgment approaches. Lower hallucination rates are better.

推理能力与幻觉率存在双向作用机制

➢ **推理增强→ → →幻觉率增加**

- **逻辑过度外推：**当模型具备强大的逻辑关联能力时，会倾向于在已知事实间建立「超合理」的虚构连接。例如，时间线延展：已知某科学家发明A技术（1990年），自动补全其在1995年获得诺贝尔奖（实际未发生）。
- **认知置信度错位：**低推理能力模型更易回答“不知道”，高推理模型会生成符合概率分布的“自信错误”答案。
- **错误前提下的正确推理：**初始假设错误，但模型基于此展开正确推理。

普通用户应对AI幻觉的三种方式

➢ **联网搜索**

”请使用联网功能“、联网功能选项

大模型	通用性测试幻觉率	事实性测试幻觉率
DeepSeekV3	2%→0%（下降2%）	29.67%→24.67%（下降5%）
DeepSeekR1	3%→0%（下降3%）	22.33%→19%（下降3%）

注：黑色为未开启联网搜索，红色为开启联网楼

普通用户应对AI幻觉的三种方式

- **双AI验证/大模型协作**

例如，利用DeepSeek生成答案后，再应用其他大模型进行审查，相互监督，交叉验证

请判断以下内容的真实性：
在《水浒传》中，李逵大闹五台山的原因主要是为了救出被囚禁的宋江。
具体情节如下：
1. **宋江被囚**：宋江因误入五台山，被寺中的和尚误认为是强盗，将其囚禁。
2. **李逵得知消息**：李逵得知宋江被囚后，心急如焚，决定前往五台山救人。
3. **大闹五台山**：李逵性格鲁莽，到达五台山后，不顾一切地冲进寺庙，与和尚们发生冲突，最终成功救出宋江。
李逵的行为反映了他对宋江的忠诚和义气，同时也展现了他冲动、鲁莽的性格特点。



Qwen2.5-Max

以上内容并不符合《水浒传》原著的情节，因此是**不真实的**。以下是具体分析：

1. 宋江被囚五台山？

在《水浒传》中，并没有宋江被囚禁于五台山的情节。宋江确实曾因杀人而逃亡，但他的主要经历包括在清风寨、江州等地的冒险，以及后来被刺配江州等情节。五台山并不是宋江的重要活动地点。

五台山在《水浒传》中与鲁智深的故事密切相关。鲁智深因打死镇关西后出家，在五台山文殊院当和尚，但他因性格豪放、不守清规戒律，多次闹事，最终被迫离开五台山。五台山的情节主要是围绕鲁智深展开的，与宋江无关。

普通用户应对AI幻觉的三种方式

➢ **提示词工程**

- **知识边界限定：通过时空维度约束降低虚构可能性（本质：约束大模型）**

1. **时间锚定法：** “基于2023年之前的公开学术文献，分步骤解释量子纠缠现象" → 规避未来时态虚构
2. **知识锚定法：** ”基于《中国药典》回答，若信息不明确请注明“暂无可靠数据支持” →限定权威来源
3. **领域限定符：** “作为临床医学专家，请列举FDA批准的5种糖尿病药物” → 添加专业身份限定
4. **置信度声明：** “如果存在不确定性，请用[推测]标签标注相关陈述“ → 减少绝对化错误断言
5. **上下文提示：** “根据《2024全球能源转型报告》（国际能源署，2024年1月发布）显示：”2030年光伏发电成本预计降至0.02美元/千瓦时，但储能技术突破仍是普及瓶颈。“请基于此数据，分析中国西部光伏基地发展的三个关键挑战，并标注每个挑战与原文结论的逻辑关联。 → 嵌入权威数据片段
6. **生成参数协同控制：** “请以temperature=0.3的严谨模式，列举2024年《柳叶刀》发表的传染病研究”

普通用户应对AI幻觉的三种方式

➢ **提示词工程**

- **对抗性提示：强制暴露推理脆弱点，用户可见潜在错误路径（本质：大模型自我审查）**

1. **植入反幻觉检测机制：** "请用以下格式回答：- 主要答案（严格基于公开可验证信息）- [反事实检查] 部分（列出可能导致此答案错误的3种假设)“
2. **预设验证条件，迫使模型交叉检查信息：** “请先回答“量子纠缠能否证明灵魂存在？”，然后从以下角度验证答案的可靠性： 1. 物理学界主流观点； 2. 近五年相关论文数量； 3. 是否存在可重复实验证据。 ”
3. **链式验证：** 请完成以下验证链：1. 陈述观点：_______2. 列出支撑该观点的三个权威数据源3. 检查每个数据源是否存在矛盾信息4. 最终结论（标注可信度等级）

幻觉高发场景

场景类别	具体场景	示例	风险等级	防护建议
知识边界模糊	开放域生成	续写未完结的经典文学作品	高	添加创作范围限制+事实性标注
	未来事件预测	预测2030年科技突破细节	极高	声明预测性质+概率分布呈现
复杂推理	多跳推理任务	追溯企业高管早期职业轨迹	高	分步验证+外部知识库检索
	数学证明延伸	要求证明未解决的数学猜想	极高	中断机制+当前研究进展说明
技术性诱发	长文本生成	小说连续章节生成	中	阶段一致性检查+人物属性维护
	多轮对话	复杂业务流程咨询	高	对话历史摘要+关键事实复核
数据引用	矛盾数据源	不同版本的实验数据引用	中	矛盾点对比+最新研究成果优先
情感驱动	安慰性回应	重症患者寻求治疗方案建议	极高	情感剥离响应+理论应用提示
特殊领域	医疗诊断	根据症状描述提供诊断建议	极高	明确非专业建议+医疗数据库
	法律咨询	解释特定法条适用范围	高	司法辖区限定+法律条文引用
	金融预测	给出具体股票买卖建议	极高	风险提示+历史回报率说明

应对AI幻觉的技术方案

- **RAG框架：**利用检索增强生成（如先搜索权威数据库，再生成答案）
- **外部知识库：**结合外部知识库，砍通用知识，强化垂直领域
- **精细训练：**针对不同任务类型进行具体的微调或强化
- **评估工具**：开发高效的自动化AI幻觉识别工具，对生成内容进行及时验证

如何应对AI幻觉？

➢ **总结**

- **三角验证法：**交叉比对多个AI回答或权威来源。
- **警惕“过度合理”：**越细节丰富的回答越需谨慎（如AI虚构论文标题与作者）。
- **理解幻觉，享受幻觉：**理解幻觉的特点和应对方法，享受幻觉带来的创意灵感

AI幻觉的创造力价值

➢ **科学发现：从“错误”到突破的范式跃迁**

- 蛋白质设计：大卫·贝克团队利用AI“错误折叠”启发新型蛋白质结构，获2024诺贝尔化学奖。
- 认为AI幻觉是“从零开始设计蛋白质”的关键
- 发表的多篇论文都将“幻觉”纳入标题当中

Article | Published: 01 December 2021

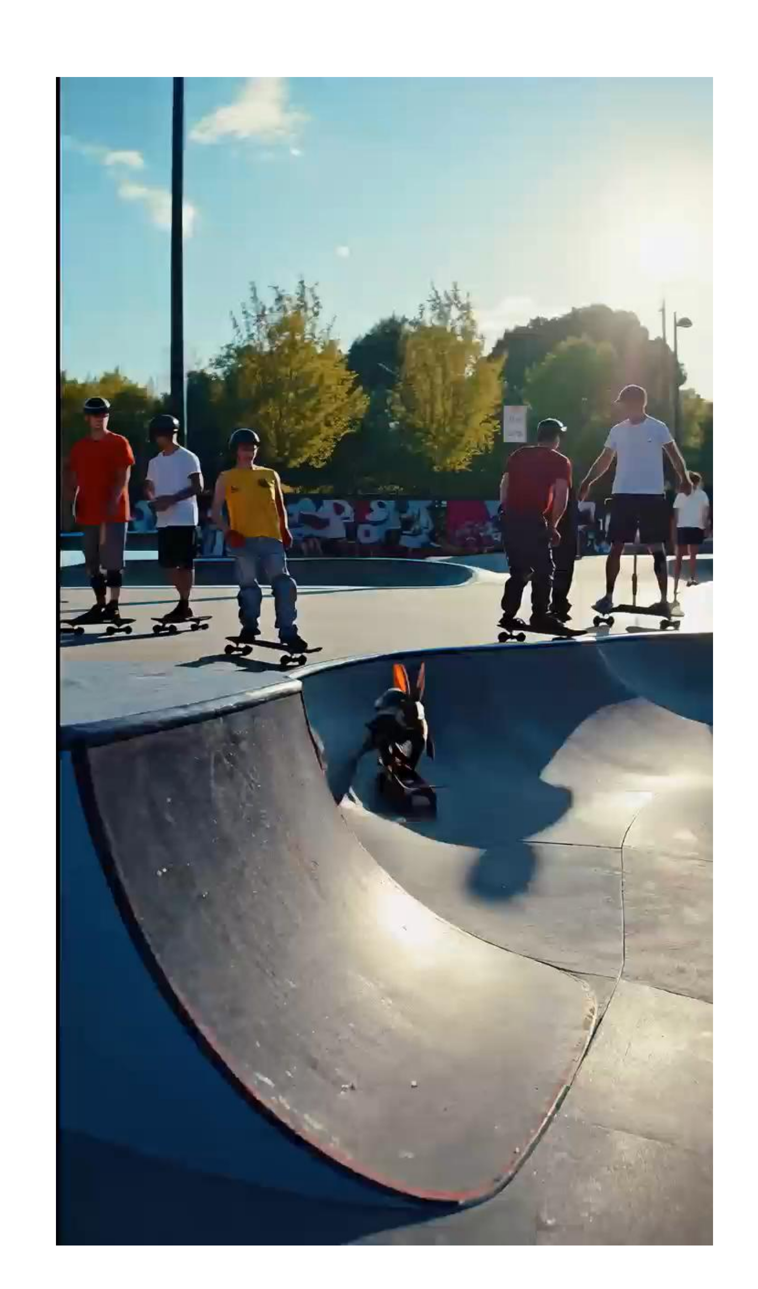
De novo protein design by deep network hallucination

Ivan Anishchenko, Samuel J. Pellock, Tamuka M. Chidyausiku, Theresa A. Ramelot, Sergey Ovchinnikov, Jingzhou Hao, Khushboo Bafna, Christoffer Norn, Alex Kang, Asim K. Bera, Frank DiMaio, Lauren Carter, Cameron M. Chow, Gaetano T. Montelione & David Baker



AI幻觉的创造力价值

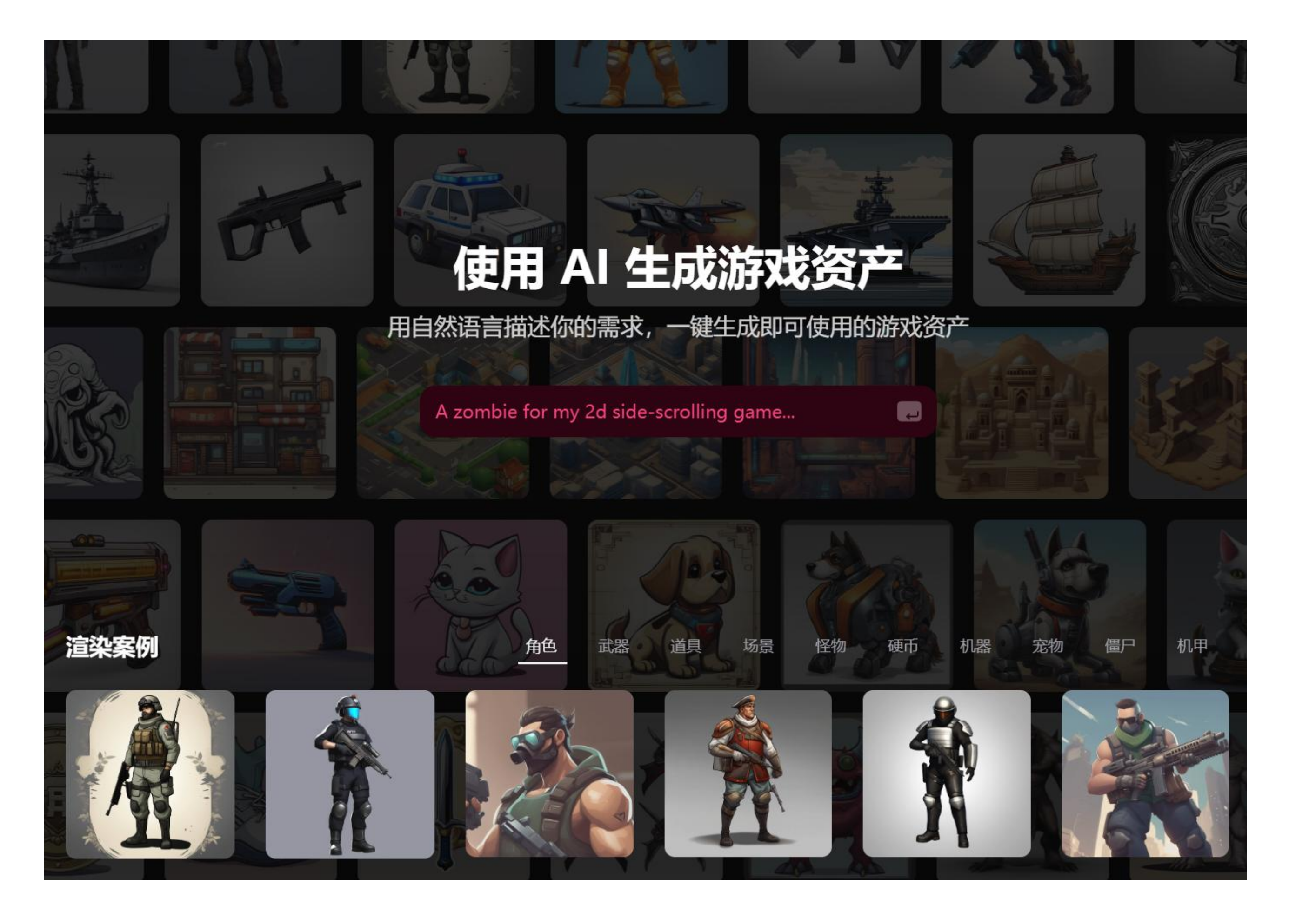
- 文艺与设计：突破人类思维定式的“超现实引擎”



AI幻觉的创造力价值

➢ **娱乐与游戏：创造新的视觉和听觉体验**

- AI生成的虚拟环境和角色设计为游戏开发人员提供了无限的可能性，增强了玩家的沉浸感和探索欲
- AI幻觉还被用于生成故事、对话和诗歌，为游戏和文学创作提供灵感



AI幻觉的创造力价值

- **技术创新：从“缺陷”到方法论的转化**

  - DeepMind团队发现，AI在图像分割任务中产生的“超现实边界”虽不符合真实场景，却意外提升了自动驾驶系统对极端天气（如浓雾、暴雨）的识别精度。

- **新型科研范式：科学界正构建“AI幻觉-实验验证-理论重构”的三阶段研究流程。**

  - 加州理工学院团队通过AI生成虚构导管设计，最终通过新型人工智能技术优化后的新设计，在实验中证实将向上游游动的细菌数量减少了100倍，形成“疯狂创意→理性筛选”的创新闭环。

AI幻觉像一面棱镜，既折射出技术的局限性，也投射出超越人类想象的可能。与其追求“绝对正确”，不如学会与AI的“想象力”共舞——因为最伟大的创新，往往诞生于理性与狂想的交界处。

——DeepSeek R1

元知（AI综述），智灵 Deep Research https://yuanzhi.zeelin.cn/#/

【免费版】



论文标题 请输入论文标题，示例：大语言模型研究综述

功能版本 基础版

样例 基础版

无数据生成报告（仅支持英文）

生成综述案例：元知（增强版）AI综述工具

《ChatGPT 与 AI 传播：规制、理解与功能整合研究》

本次研究选取中国学术期刊网络出版总库 CNKI 和美国科学情报研究所(Institute for Scientific Information，ISI) 的 Web of Science（WOS）数据库（时间跨度选取为 2023—2024 年）作为切入点，分别获取中英文有效文献 20 篇、17 篇。

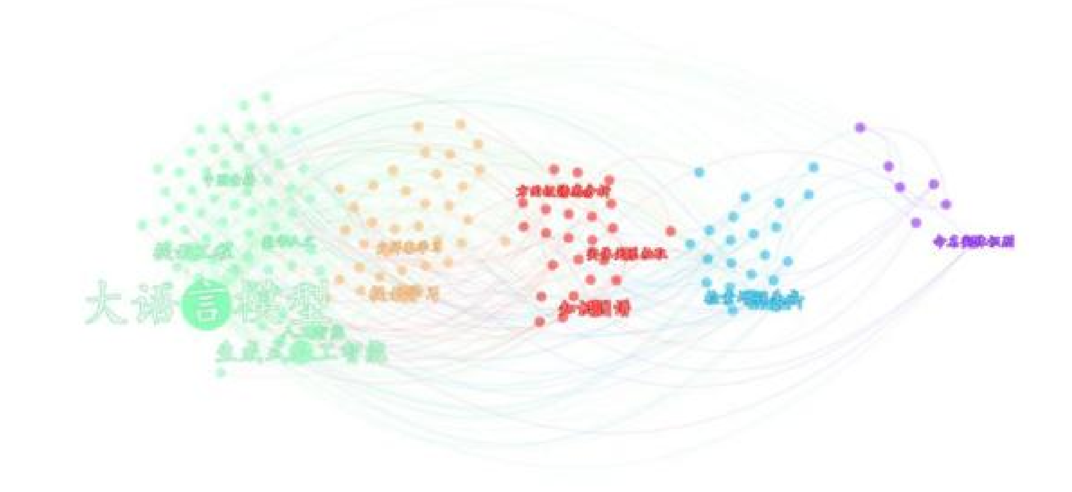


图 1研究主题关键词共现聚类图谱

大语言模型传播偏向规制与风险治理：以 ChatGPT 为例

一、研究现状

1. 大语言模型传播规制研究层面

重点关注大语言模型在不同领域的应用与挑战。郑春萍等（2024）提出，人工智能在语言教学领域的应用促使自然语言处理、机器学习等前沿技术方法得到广泛应用，从而促进学习者的知识获取与技能习得，对核心素养塑造、学习心理分析及策略行为发展产生积极影响[7]。孟旭阳等（2024）提出，通过深度学习技术优化文本摘要模型，并利用大语言模型实现结构化综述生成，有效提升了学术文献的知识化服务水平，使得学术信息处理效率显著提高[9]。刘邦奇等（2024）提出，生成式人工智能的显著突破及其在教育领域的深度应用，将促进教育主体关系转变、环境智能升级、资源供给创新等变革，进而助力人类教育与学习形态的重塑[10]。苏君阳等（2024）认为，大语言模型在学术研究中的应用虽带来原创性、知识管理与应用认同等价值，但结构性与能动性局限易造成研究信效度难以认定、人机角色责任划分不清，进而产生学术伦理不端与研究者关键技能退化的风险[11]。于千雯等（2023）认为，AIGC 技术在学术论文生产中的应用能有效协助作者和编辑，但同时也带来了学术道德、技术局限和版权合规等问题，使得学术期刊编辑面临新的机遇与挑战，需要从应用、治理和素养提升三方面寻求发展路径[12]。徐敬宏等（2024）提出，大语言模型的应用在学术出版中提高了效率和智能化水平，但同时也引发了著作权侵犯、学术垃圾、信息安全隐患等问题，因此学术出版机构需加强人工监管和规范使用[13]。韩筠（2023）提出，数字平台建设和应用推动了高等教育教学创新，通过引入大语言模型等人工智能新技术，优化平台功能，升级技术应用，生成新的教学服务模式，从而构建泛在学习环境下的智慧教育生态，使得教学创新开辟新领域，产生显著的教育变革效应[19]。吴冠军（2023）认为，以 ChatGPT 为代表的大语言模型虽展现出通用智能，却频发错误，这从技术政治学视角出发，揭示了其错误生成与意识形态偏见之间的因果关联，进而强调在人工智能时代，意识形态批判性分析的重要性[20]。Tanksale V（2023）提出，大语言模型在 Web3D 应用中的集成能够显著促进内容生成、自然语言交互、个性化及知识整合，但同时也带来了伦理挑战，并为此领域未来的研究方向提供了新的视角[21]。Pester A（2024）提出，大型语言模型在自然语言处理领域的突破性进展，成功应用于沉浸式学习环境，这不仅符合教学原则，还显著提升了现有教育系统的有效性[26]。Bonnechere B（2024）认为，大型语言模型的运用能够显著提升康复治疗过程的数据整合与决策，通过解决数据偏见、语境理解及伦理问题，促进康复领域的进步与优化[27]。Hobensack M（2024）认为，尽管大型语言模型在护理实践、教育和研究中的应用存在显著机遇，但其使用和采纳引发了诸如偏见、误用和剽窃等伦理问题，从而造成了对建立评估、评价、标准和指南的持续需求，以确保其适当、准确和安全的使用[30]。Chen ZY（2024）认为，随着大型语言模型（LLM）的快速发展，其在自然语言处理领域的贡献显

人机快生之元境平台

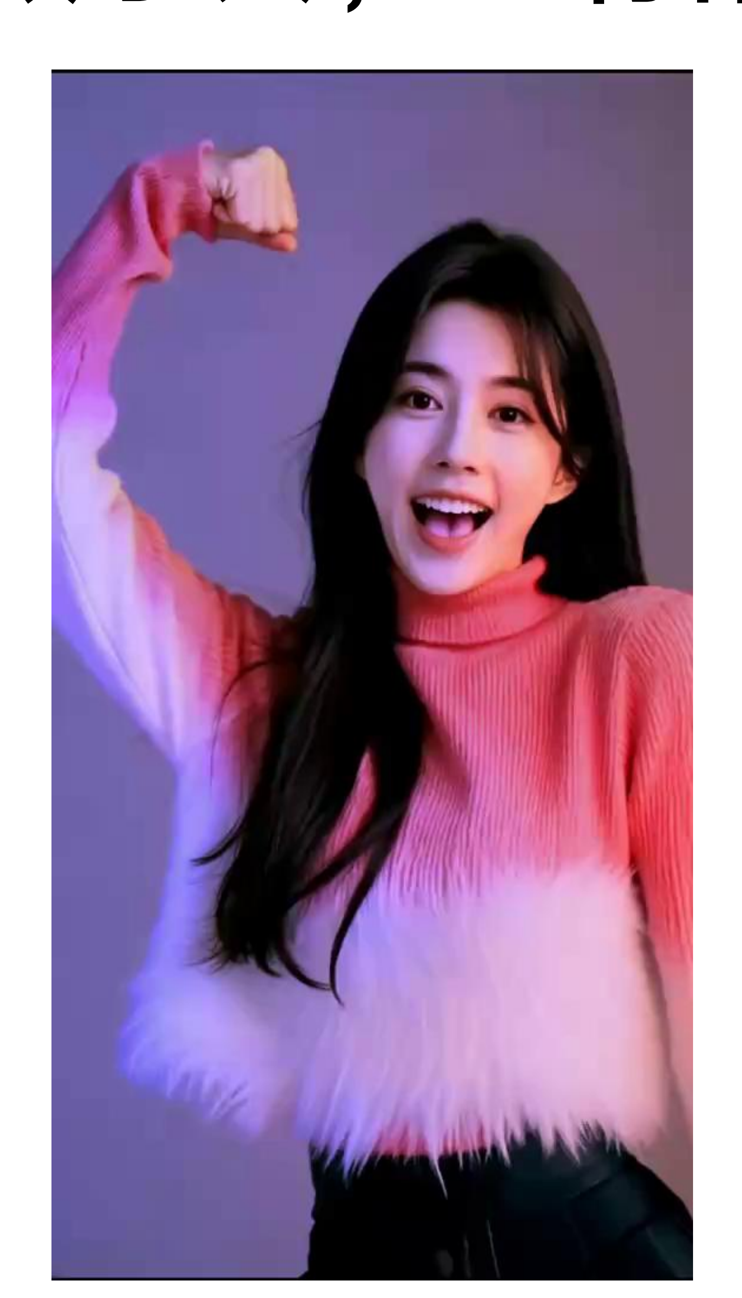
提示词：懒猫街舞送祝福，极简画风迎新春。



- 入口网址：https://yuanjing.zeelin.cn/#/case
- 访问二维码：



自制数字人，已制作200多个



自研人形机器人



自制AI微短剧，已制作300多段





谢谢！





观看完整视频讲解，请关注“新媒沈阳”视频号



获取更多PPT资源，请关注“清新研究”二维码